DOCUMENT RESUME

ED 222 565 TM 820 723

AUTHOR Roeber, Edward D.
TITLE Using Performance Tests in State Assessment--It's Real and It's Feasible.
PUB DATE May 82
NOTE 9p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Behavioral Objectives; *Educational Assessment; Elementary Secondary Education; *Performance Tests; *State Programs; Student Evaluation; Test Construction; *Test Format; *Testing Programs; Test Use
IDENTIFIERS Michigan; *Michigan Educational Assessment Program

ABSTRACT

Each year the Michigan Educational Assessment Program (MEAP) tests all 4th, 7th, and 10th grade students in reading and mathematics and one or more subject areas. Because MEAP has lost funding for test development, experienced assessment staff and a volunteer team of local educators and college and university specialists develop, administer, and score the MEAP tests. It is important to both instructional and assessment specialists to develop measures to match the skills being tested. Some performance objectives can be tested with multiple choice items, but some skill areas such as art, career development, health, music, and science can only be measured with open-ended short-answer or essay formats, classroom observation, or individual or group performance tests. The performance data yielded can be valuable for curriculum review and planning at local and state levels. Test development activities are conducted several times for each subject area to determine appropriate item types for each objective. The use of a variety of item types provides more valid information to teachers and content specialists. (CM)

USING PERFORMANCE TESTS IN STATE ASSESSMENT--IT'S REAL AND IT'S FEASIBLE

Edward D. Roeber

INTRODUCTION

Each year, all fourth, seventh and tenth graders in Michigan take part in the Michigan Educational Assessment Program (MEAP). Students take a mathematics and reading test comprised of multiple choice items. Also each year, samples of students at the same grade levels are tested in one or more other subject areas--Art, Career Development, Health, Music, Physical Education, Science, Social Studies, Speaking and Listening, and Writing. Performance objectives have been written in each of these areas; many of the objectives are tested with multiple choice items. However, other objectives require the use of open-end formats (short answer or essay), classroom observation, or even individual or group performance tests.

MEAP has been able to develop and use measures that are congruent with the underlying skills. It is feasible to develop such tests, administer them in a large-scale assessment program, and score student responses. Although the tests are more difficult to write and to use, they yield valuable data for curriculum review and planning at both the local and state levels. The purpose of the paper is to suggest some ways such tests can be used in large-scale assessment programs. While some of these skills could have been "converted" to multiple choice test items, students' performance on these important (even critical) skills could not be accurately determined. It is important to both instructional and assessment specialists that the measurement used match the skills to be tested.

SKILLS REQUIRING NON-TRADITIONAL MEASURES

For each area in which performance objectives have been written (see above), there are some skills that curriculum specialists inevitably include that can only be measured with open-end or performance items. Listed below are just a few of the ones which could be found in Michigan.

SUBJECT AREA	SKILLS
Art	Draw, paint--participate in art activities
Career Development	Apply for a job, interview for a job, participate in group discussions
Health	Brush teeth properly, demonstrate first aid
Mathematics	Measure weight, length, volume
Music	Sing alone and with groups, dance, perform on musical instruments
Physical Education	Running endurance, throw and catch a ball, strike a ball
Reading	Use reference materials
Science	Conduct simple experiments, graph data
Social Studies	Participate in group discussions, make decisions, present information
Speaking and Listening	Speak in public, analyze conversations, communicate non-verbally
Writing	Write an essay or letter

As planning for the assessment of each subject area begins, skills requiring multiple-choice items, open-end items, or individual or group performance are identified. In some areas, subject-matter specialists will write more objectives requiring one type of measure, than another. However, at least some open-end and performance items are called for in each of the subject areas.

DEVELOPING THE TESTS

MEAP has lost the funding to carry out test development projects via contract. Although this approach was used ealier (1972-1977), it simply is no longer feasible. However, there are MEAP staff with test development experience and there are interested local educators and college and university specialists often eager to see "their" subject area assessed. At the heart of the test development and test administration is this core group of volunteers. Without them, much of the work would not be possible.

The tests which measure the Michigan objectives are developed by the volunteer team working with the MEAP staff. In order to build a liaison with the appropriate subject-matter organization, a team leader is jointly appointed by the organization and MEAP. Ideally, the team leader would be a well-respected college or university specialist with experience working in schools. The team is filled in by the team leader and MEAP to represent differing instructional philosophies (if any), levels of work assignment and regions of the state.

The MEAP staff, Department instructional staff, and team leader first meet to determine the appropriate item types for each objective. The team is trained by MEAP and works under the supervision of both MEAP and the team leader. If necessary, graduate assistants may be used where needed (e.g., item editing and item revisions). Once trained in item writing, the team works to write the items needed. Where possible, other items that are already available will be used--the NAEP items for example. Even in instances where a particular item will not measure a Michigan objective, the underlying measurement techniques can be adapted to fit. For example, the NAEP techniques for assessing music performance were used, although new items were written. Borrowing is particularly sensible for open-end or performance items, where developing an administerable and scorable item is very time-consuming.

After the items are written, the team leader (and assistant) and Department staff edit the test items, package the items into tests, conduct a subject-matter specialist review, and prepare for tryouts. After tryouts, the items are again reviewed by the specialists, and finally, the team leader and Department staff select the items to be used and finalize the test packages.

The above-mentioned activities have been conducted several times across the various subject areas assessed in MEAP. About a year is required to develop the tests and costs range from $1,000 to about $10,000, depending on complexity of the area, specialized materials needed, number of team members and number of team meetings. While volunteers are sought for the team (to work without pay), expenses are paid. A small honorarium is paid to the team leader and assistants in recognition of the extent of their contributions. If conceived and carried out as a cooperative effort, though, tests can be developed at very little expense.

ADMINISTERING THE TESTS

While the volunteers can develop the open-end or performance tests, how can the costs of administering the tests be reduced? Can such items be feasibly used in an assessment program? Is it valuable to use such tests--that is, can such tests provide useful information?

The non-multiple choice tests fall into two types--those which can be administered to students in group settings by classroom teachers and those that are individually administered by a trained test administrator. The former category includes some special types of items, including writing essays, music or listening items presented via cassette tapes and so forth. Although they are not traditional items, they can be administered in much the same way as the multiple-choice tests.

However, the individual performance tests are another matter. It is not realistic to expect classroom teachers to be able to administer a test such as a music performance test individually to students for several reasons. First, many schools may not have a teacher in the content area to be tested. Second, even if there is a teacher, he or she will need to be trained to administer such tests, since no one may be available to cover their classes while they administer the tests. However, it is difficult to provide this training. Third, use of regular teachers is disruptive to the instructional schedule for all students in the class. Again, the use of volunteers is a cost-effective solution to this problem.

Because volunteers, recruited through the subject-matter organization, were used to write the items, the subject-matter organization and team leader should be used to locate volunteers to administer the tests. Often the team leader becomes the individual who leads the test administration effort as well. The effort to recruit test administrators begins with the universities offering graduate level training in the content area. In fact, this experience has proven so valuable for students that some universities strongly promote the test administration as a training activity for the graduate students.

Other volunteers for the test administration may come from unemployed teachers or from content area supervisors in local or intermediate districts. For this reason, the test sample is drawn several months in advance, so that the schools in which testing will take place can be notified and the administration team leader can see if anyone in the district has the time necessary to give the tests. This is particularly valuable in two instances: 1) urban areas where a relatively large number of students are to be tested and 2) in rurally-isolated districts, where travel costs could be high if an outsider had to come in to do the testing.

The typical test project involved one test package at each grade tested. The test usually takes between thirty to sixty minutes to administer. About twenty-five test administrators are needed. Each will devote about 3 to 4 days to the task during the four-week assessment period. Anywhere from thirty to over one hundred different schools may be involved, although only a small sample of students (and alternatives) are actually tested. Test administrators are reimbursed for their expenses, but are not paid an honorarium or fee to do the testing. Expenses of test administration include preparation of the test booklets, testing manuals, tapes or other special materials or handouts, as well as the test administrators' expenses. The testing materials costs have ranged from under $2,000 to over $20,000; the test administrators' expenses range from about $2,000 to $4,000.

Local districts that are interested in doing their own testing can either be trained in doing so by the MEAP staff, or can use one of the trained test administrators. Obviously, local districts would need to make their own financial arrangements, particularly in the latter case. However, this does provide a mechanism for extending testing to local schools.

SCORING THE TESTS

The non-multiple-choice items, whether group or individually administered, often require specialized scoring. Essay questions and other written responses require the development of open-end scoring guides. Then, the guides need to be applied to all of the students' responses. Music or Speaking tests require the scoring of cassette tapes, which is a similar problem to written essays but presented on a different media. Other areas like Physical Education require tests administrators to score student responses as they occur, so the test administrators that conduct the scoring need to trained in advance.

In each case, the scoring guides begin at the item writing stage, are more fully refined before tryouts, and are carefully reviewed after tryouts. If test administrators need to learn how to score students' responses, this scoring is built into the test administration training. If not, responses are gathered and another volunteer group actually conducts the scoring. Several groups are used to solicit volunteers: 1) item writer volunteers, 2) test administrators, 3) local districts in which testing took place and 4) other local districts that conducted testing. Group size and length of scoring obviously depend on the number and complexity of responses to be scored, but again costs are limited to scorer expenses.

A major advantage of using interested local educators in the scoring is that they learn the scoring techniques and can apply them locally. This provides a group of trained scorers who can assist in the application of the tests locally, both in the conduct of local assessments but also in classroom instruction where appropriate.

IS PERFORMANCE TESTING USEFUL?

By this point, the reader hopefully has concluded that performance testing can be a feasible part of a Large-Scale Assessment Program. But, is it worth it? Why go through all of this to collect the data? The answer is that it indeed is worth it. While few in number, these skills represent the "heart" of these content areas and are what both professional and lay people see are the end products of instruction in the area. If students cannot perform these skills, then there are definite problems. While the multiple-choice items can suggest the existence or causes of problems, they cannot demonstrate the problems as graphically as the performance data does.

In addition, by testing these skills, the entire content assessment and by implication, the assessment program itself takes on added (and deserved) aura of content validity. The data from the entire assessment, even the non-performance items, can be trusted to a greater extent. Appearing more valid, the information is more likely to be put to use by teacher and content specialists. The costs are low, considerable effort is required, but the payoffs are greater.

SUMMARY

It is possible to use non-multiple choice tests in a large-scale assessment program. Testing can be carried out at very little cost and it is feasible to use these items. Because of the importance of the skills tested, the data the tests yield is read and can be more easily translated into curricular suggestions by teachers and specialists. It also helps to "bridge the gap" between curriculum specialists and test experts, which carried over into greater acceptance of all of the assessment information. For all of these reasons, the use of performance tests is feasible--and it is real.

EDR:se:5/17/82